network *comparison*

introduction to *network analysis* (*ina*)

Lovro Šubelj
University of Ljubljana
spring 2023/24

comparison *overview*

- network *comparison by isomorphism* is *NP problem*
- *exact comparison* requires *exponentially many* properties

- comparison using selected *graph edit distance*
- comparison by *network fragments* [MIK+04, Prž07, ARS15]
- comparison by *network distances* [SCDG+17, BB19]

- direct comparison of *individual metrics* [WS98, BA99, New02]
- statistical comparison over *multiple metrics* [ŠFB14, ŠBB+15]

comparison *metrics*

CONSISTENCY OF CITATION AND COLLABORATION TOPOLOGY OF BIBLIOGRAPHIC DATABASES

Šubelj, L., Fiala, D. & Bajec, M. **Network-based statistical comparison of citation topology of bibliographic databases**. *Scientific Reports* **4**, 6496 (2014).
Šubelj, L., Bajec, M., Boshkoska, B., Kastrin, A. & Levnajić, Z. **Quantifying the consistency of scientific databases**. *PLoS ONE* **10**(5), e0127390 (2015).

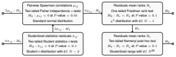
Corresponding author: lovro.subelj@fri.uni-lj.si

NETWORKS OF BIBLIOGRAPHIC DATABASES

Citation and collaboration networks extracted from bibliographic databases. These are: **(WoS)** the Computer Science category of Web of Science until 2014 (979k papers); **(APS)** the American Physical Society publications until 2010 (450k papers); **(PubMed)** the PubMed Central Collection open access publications until 2014 (5.9M papers); **(DBLP)** the DBLP Computer Science Bibliography until 2014 (2.7M papers); **(arXiv)** the High Energy Physics Theory category of arXiv between 1992 and 2003 (28k papers); **(CiteSeer)** web publications parsed by the CiteSeer service (723k papers); **(Cora)** McCallum's Cora database collected from the web in 1998 (196k papers); and **(HistCite)** Lederberg's bibliography produced by the Algorithmic Historiography (9k papers).

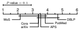
NETWORK COMPARISON METHODOLOGY

Methodology of network-based statistical comparison of bibliographic databases. Networks representing bibliographic databases are compared through 21 graph statistics. We compute externally studentized statistics residuals that measure the consistency of each database with the rest. Statistically significant inconsistencies in individual statistics are revealed by independent Student *t*-tests. We select a subset of statistics whose pairwise independence is verified using Fisher *z*-transformation. Friedman rank test confirms that databases display significant inconsistencies in the selected statistics, while the databases with no significant differences are revealed by Nemenyi post-hoc test.

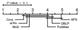


COMPARISON OF BIBLIOGRAPHIC NETWORKS

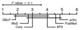
Statistical comparison of bibliographic databases through statistics of networks. Panel **(A)** shows the critical difference diagram of Nemenyi test for paper citation networks P→P, panel **(B)** for author citation networks A↔A and panel **(C)** for author collaboration networks A—A (no additional author name disambiguation has been made). The critical diagrams illustrate the overall ranking of the databases, where those connected by a thick line show no statistically significant inconsistencies at *P*-value = 0.1.



A P→P



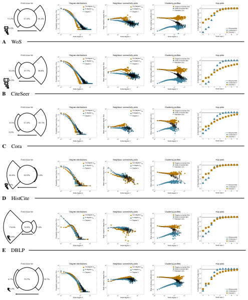
B A↔A



C A—A

PROFILE OF PAPER CITATION NETWORKS

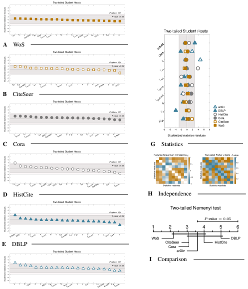
Distributions, diagrams, plots of paper citation networks extracted from bibliographic databases. Panels **(A–F)** show (from left to right): the field bow-tie decompositions, where the arrows illustrate the direction of the links and the areas of diagrams are proportional to the number of nodes with zero out-degree, non-zero degree and zero in-degree, respectively; the degree, in-degree and out-degree distributions $P(k)$, $P(k_{in})$ and $P(k_{out})$, respectively; the degree mixing by the corresponding neighbour connectivity plots $N(k)$, $N(k_{in})$ and $N(k_{out})$; the clustering profiles of the standard, degree-corrected and delta-corrected coefficients $C(k)$, $D(k)$ and $B(k)$, respectively; and the hop plots for the directed and undirected 90-percentile effective diameters d and d´, respectively.



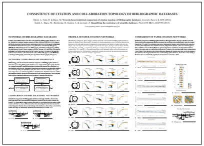
A WoS
B CiteSeer
C Cora
D HistCite
E DBLP
F arXiv

COMPARISON OF PAPER CITATION NETWORKS

Statistical comparison of bibliographic databases through statistics of paper citation networks. Panels **(A–F)** show studentized statistics residuals that are listed in decreasing order, while the shaded regions are 95% and 99% confidence intervals of independent Student *t*-tests (labelled with respective *P*-values). Panel **(G)** shows the residuals of merely independent statistics, where the shaded region is 95% confidence interval. Panel **(H)** shows pairwise Spearman correlations of independent statistics listed in the same order as in panel **(G)** (left) and the *P*-values of the corresponding Fisher independence *z*-tests (right). Panel **(I)** shows the critical difference diagram of Nemenyi post-hoc test for the independent statistics. The diagram illustrates the overall ranking of the databases, where those connected by a thick line show no statistically significant inconsistencies at *P*-value = 0.05.



A WoS
B CiteSeer
C Cora
D HistCite
E DBLP
F arXiv
G Statistics
H Independence
I Comparison



↓ *statistical comparison* of N *networks* over K *metrics*

1. x_{ij} is *value* of jth *metric* for ith *network* and $\widetilde{x}_{ij}$ its *residual*

$$\widetilde{x}_{ij} = \frac{x_{ij} - \widetilde{\mu}_{ij}}{\widetilde{\sigma}_{ij}\sqrt{1-\frac{1}{N}}} \qquad \widetilde{\mu}_{ij} = \frac{1}{N-1}\sum_{k\neq i} x_{kj} \qquad \widetilde{\sigma}_{ij} = \sqrt{\frac{1}{N-2}\sum_{k\neq i}(x_{kj} - \widetilde{\mu}_{ij})^2} \qquad \widetilde{x}_{ij} \sim t(N-2)$$

2. R_{ij} is *rank* of ith *network* for jth *independent metric*

$$R_{ij} = rank\ of\ |\widetilde{x}_{ij}| \qquad R_{ij} \in \{1, \ldots, N\}$$

R_i is *mean rank* of ith *network* over K *independent metrics*

$$R_i = \frac{1}{K}\sum_j R_{ij} \qquad \frac{12K}{N(N+1)}\left(\sum_i R_i^2 - \frac{N(N+1)^2}{4}\right) \sim \chi^2(N-1)$$

3. $|R_i - R_j|$ *statistically significant* when above *critical difference* $q\sqrt{\frac{N(N-1)}{6K}}$

comparison *references*

David Aparício, Pedro Ribeiro, and Fernando Silva.
Network comparison using directed graphlets.
e-print arXiv:151101964v1, 2015.

A.-L. Barabási and R. Albert.
Emergence of scaling in random networks.
Science, 286(5439):509–512, 1999.

A.-L. Barabási.
Network Science.
Cambridge University Press, Cambridge, 2016.

James P. Bagrow and Erik M. Bollt.
An information-theoretic, all-scales approach to comparing networks.
Appl. Netw. Sci., 4:45, 2019.

Wouter de Nooy, Andrej Mrvar, and Vladimir Batagelj.
Exploratory Social Network Analysis with Pajek: Expanded and Revised Second Edition.
Cambridge University Press, Cambridge, 2011.

David Easley and Jon Kleinberg.
Networks, Crowds, and Markets: Reasoning About a Highly Connected World.
Cambridge University Press, Cambridge, 2010.

Ernesto Estrada and Philip A. Knight.
A First Course in Network Theory.
Oxford University Press, 2015.

comparison *references*

Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon.
Superfamilies of evolved and designed networks.
Science, 303(5663):1538–1542, 2004.

M. E. J. Newman.
Assortative mixing in networks.
Phys. Rev. Lett., 89(20):208701, 2002.

Mark E. J. Newman.
Networks.
Oxford University Press, Oxford, 2nd edition, 2018.

Nataša Pržulj.
Biological network comparison using graphlet degree distribution.
Bioinformatics, 23(2):e177–e183, 2007.

Lovro Šubelj, Marko Bajec, Biljana Mileva Boshkoska, Andrej Kastrin, and Zoran Levnajić.
Quantifying the consistency of scientific databases.
PLoS ONE, 10(5):e0127390, 2015.

Tiago A. Schieber, Laura Carpi, Albert Díaz-Guilera, Panos M. Pardalos, Cristina Masoller, and Martín G. Ravetti.
Quantification of network structural dissimilarities.
Nat. Commun., 8:13928, 2017.

Lovro Šubelj, Dalibor Fiala, and Marko Bajec.
Network-based statistical comparison of citation topology of bibliographic databases.
Sci. Rep., 4:6496, 2014.

comparison *references*

D. J. Watts and S. H. Strogatz.
Collective dynamics of 'small-world' networks.
Nature, 393(6684):440–442, 1998.